A Survey on Efficient Inference for Large Language Models

Year: 2024 | Citations: 162 | Authors: Zixuan Zhou, Xuefei Ning, Ke Hong

---

Abstract

Large Language Models (LLMs) have attracted extensive attention due to their remarkable performance across various tasks. However, the substantial computational and memory requirements of LLM inference pose challenges for deployment in resource-constrained scenarios. Efforts within the field have been directed towards developing techniques aimed at enhancing the efficiency of LLM inference. This paper presents a comprehensive survey of the existing literature on efficient LLM inference. We start by analyzing the primary causes of the inefficient LLM inference, i.e., the large model size, the quadratic-complexity attention operation, and the auto-regressive decoding approach. Then, we introduce a comprehensive taxonomy that organizes the current literature into data-level, model-level, and system-level optimization. Moreover, the paper includes comparative experiments on representative methods within critical sub-fields to provide quantitative insights. Last but not least, we provide some knowledge summary and discuss future research directions.